CIS II Final Report: Detection and Guidance of K-Wire Placement in Pelvic Trauma Surgery

Students: Irina Bataeva, Kinjal Shah
Mentor: Dr. Ali Uneri

May 13, 2020

Abstract

Minimally invasive pelvic fixation surgery involves insertion of cannulated screws, guided by Kirschner wires (K-wires), to join fractured bone fragments. The manual insertion of K-wires, however, has high injury rate, takes significant time, and requires multiple intraoperative fluoroscopes. Robotic guidance has the potential to improve the safety and efficiency of this procedure. Conventional marker-based tracking solutions, however, fail to capture potential deflections of the K-wires, which are otherwise conspicuous in radiographic imaging. We utilize deep learning tools for identifying K-wires on the pelvic x-rays, which can then be used as an input for 2D to 3D registration for K-wire localization and guidance. We present here a transfer learning approach leveraging a dataset generation workflow which combines real pelvic radiographs with simulated K-wires. We developed a model using the U-Net architecture and evaluated performance on a test dataset of 33 real patient images with K-wires. Thorough analysis with selected figures of merit has been performed and potential for future work has been identified.

1 Introduction

Pelvic fractures occur across all demographics and are estimated to be 2–8% of all fractures in the United States [1]. Cases in younger individuals are mostly due to high energy trauma such as traffic accidents, while in older demographics the primary cause is injury due to falls. Osteoporosis often causes bones to get weaker with age, and the incidence of pelvic fractures are expected to increase with the aging population of the United States [2]. As a result, pelvic fracture treatment is a growing market and is estimated to reach $1.8 billion by 2025 [3]. Although some cases can be treated with external stabilization, unstable pelvic fractures require surgical stabilization. Minimally invasive surgical (MIS) procedures are becoming an increasingly popular choice for some surgeons as they decrease blood loss, risk of infection, and recovery times for patients with unstable pelvic fractures compared to the open surgery approaches [4]. Percutaneous insertion of K-wires is challenging even for experienced surgeons due to the complex morphology of the pelvis and K-wires are flexible and thus bend inside the

patient. Surgeons, therefore, cannot use conventional marker-based navigation methods that assume a rigid guided object and are not able to accurately guide the trajectory of the K-wire. To guide the K-wire, surgeons rely on intraopertive fluoroscopy, taking images throughout the insertion of the K-wire to visualize its position relative to patient's anatomy. As a consequence, patients often get exposed up to 2 minutes of radiation per screw [5]. Despite the extensive imaging, however, the surgeons still struggle to achieve the permissible accuracy for the K-wire's trajectory of 1 mm translational error and 5° rotational error [6]. Since the K-wire is used to place stabilization screws, screw malposition is a common complication of pelvic fracture stabilization surgeries: 20–30% screw placements are rated as suboptimal [7] and 6% breach the cortical bone [8]. Screw malposition is a serious issue as it can result in neurological and vascular injuries, require longer surgical times, and lead to long-term pelvic instability [9].

2 Background and Motivation

The I-STAR lab is working on introducing surgical navigation protocol for pelvic fixation surgeries. A key component is rapid 3D localization of K-wires from X-ray images. Various approaches have been attempted to achieve identification with high speed and accuracy. One of them was the deformable known-component registration, in which a B-spline-based mesh cylinder was transformed to match the projection on the fluoroscopic image. A registration error of 2.1 ± 0.3 mm and $0.8^\circ \pm 1.4^\circ$ tip orientation was achieved. Another approach involved performing a multi-resolution pyramid technique to detect the K-wire, which was faster than the B-spline based approach, but still limited; it took 2 minutes to get 1.1 ± 1.6 mm accuracy [5]. While both approaches discussed above are novel and achieve high accuracy, they do not yet meet the speed and accuracy requirements necessary to support true surgical navigation with reduced injury.

To achieve this improved accuracy and speed, researchers have been exploring the application of deep learning tools to the K-wire and guide-wire detection problem. One of the primary limiting factors in these solutions are the availability of the large training datasets necessary to build such models. These training sets are usually limited in number and require an arduous, manual annotation process. Therefore, there has been a large interest in developing simulated datasets that are able to emulate real images [10, 11].

Our work proposes a transfer learning approach using an entirely simulated dataset of arbitrarily large sizes. We have created an adaptable dataset generation protocol and developed baseline metrics for K-wire identification tasks upon which additional analysis can be performed. We hope this work will improve K-wire insertion accuracy to reduce injury while limiting radiation exposure.

3 Methods

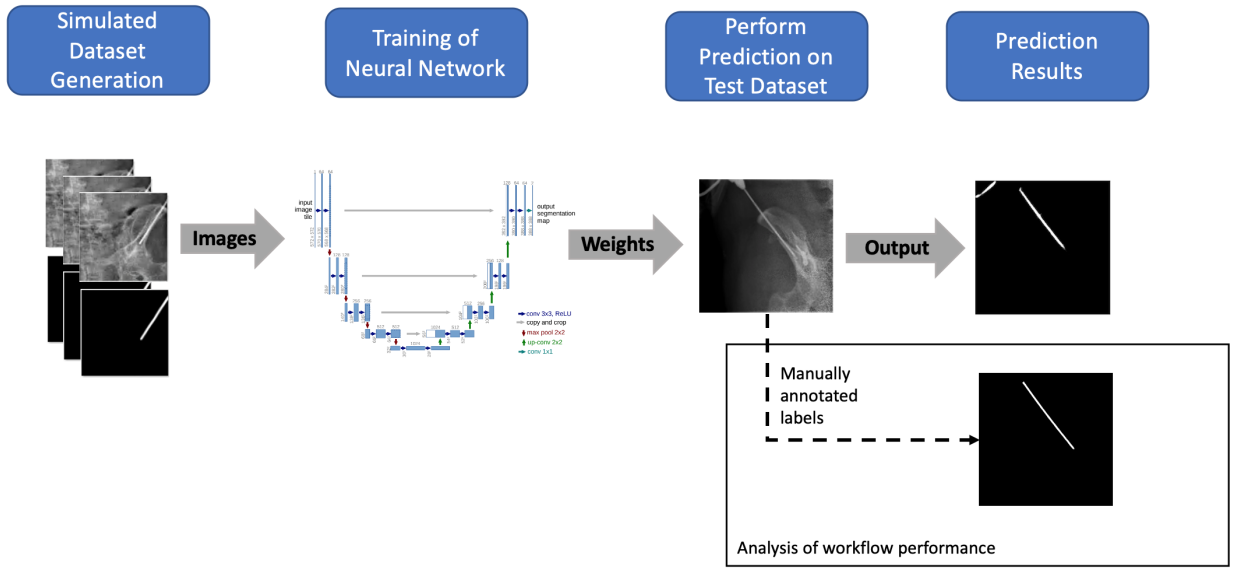

Figure 1: Flowchart describing the end-to-end approach from training to prediction. The simulated dataset is used to train a convolutional neural network (CNN) and obtain model weights. The trained weights are used to infer predictions on a test dataset consisting of real patient images from an IRB-approved dataset. The prediction results are compared to manually annotated labels of the test dataset for performance analysis.

3.1 *Training* Dataset Generation

We propose an approach for generating training data from a small number of real fluoroscopic images of the pelvis via data augmentation of anatomical structures and simulation of surgical instruments. Similar data augmentation workflows have been performed and shown to be successful in studies of guide-wire detection for alternative procedures [11]. Leveraging available medical images and increasing the training dataset size through data augmentation has been a promising technique for improving the application of deep learning methods to medical image analysis [12].

Real fluoroscopic pelvic images in DICOM format were provided by the I-STAR lab faculty consisting of 3D scans of different regions-of-interest (ROI) of 2 cadaver pelves. A K-wire model and generator, based on prior work, was also provided by I-STAR lab faculty. From the 3D scans, we selected unique images with anteroposterior (AP) and AP-like views (corresponding to C-arm gantry angles ranging $-30°$–$30°$). We constrained our initial analysis to these AP-like views, as these are commonly used for K-wire navigation as lateral views often have lower contrast and higher noise (due to increased attenuation).

Data Augmentation. To increase the statistical variation in the background images, we have applied common data augmentation techniques consisting of random rotations ($\pm 2.5°$), translations ($\pm$ 1%), scaling both in x and y-axes ($\pm$ 20%), and addition of white Gaussian noise with a standard deviation of 0.3. The images were converted from arbitrary detector

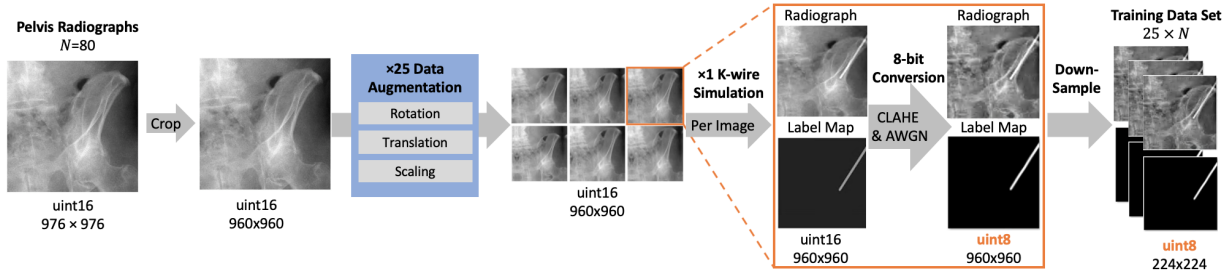

Figure 2: Process diagram for training dataset generation for a target training population of 2,000 images. Multiple ($N = 80$) fluoroscopic images extracted from 3D cone-beam CT scans of cadaveric pelvic specimens undergo data augmentation ($\times 25$), followed by simulated K-wire projections on each augmented image to produce $25 \times N$ images and K-wire labels for use in training an artificial neural network for K-wire detection.

units to line integrals of attenuation coefficient and log-corrected. Each image was augmented 25 times with random selection of the augmentation parameters, resulting in a $\times 25$ increase in background image count.

K-Wire Simulation. The B-spline shape of the K-wire was projected onto each of augmented pelvic images. The following variations in the K-wires were applied: radius (1.3–1.9 mm), length (20–30 cm), tip length (1–4 mm) to account for variations in K-wires used for pelvic trauma surgeries. Translation, rotation, and curvature were also randomized. Attenuation of the projection was set to be be randomly selected from a range between $0.08 and 0.2 mm^{-1}$. The K-wire was projected to both the pelvic image and a blank image to create a pixel-by-pixel ground truth for the K-wire's location on the pelvic image. We created datasets with both rigid only K-wires as well as bent K-wires to evaluate performance. See Figs. 3 and 4 for examples of the images from the rigid and deformed datasets.

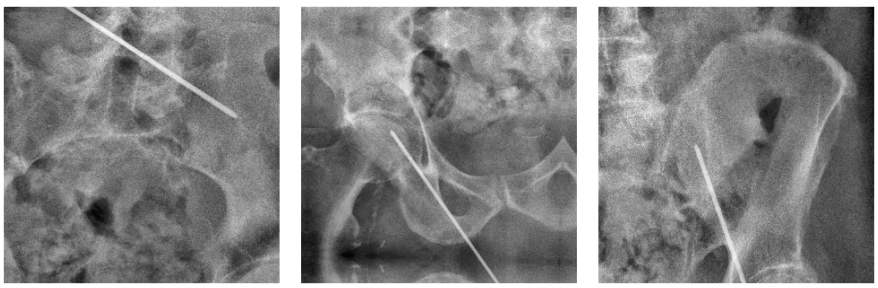

Figure 3: Example images from the generated training dataset with rigid simulated K-wires. Simulated B-spline based K-wire models with 2 control points (i.e., line) were projected at varying poses on the augmented images of anatomy.

Image Post-Processing for Input to Network. In order to prepare the images for input into the network architecture, we needed to perform a 16- to 8-bit conversion. In order to minimize information lost during this conversion we applied a contrast limited adaptive histogram equalization on the images. The image was then converted into an 8-bit image. The final step was to downsample the images using bilinear interpolation for radiograph and

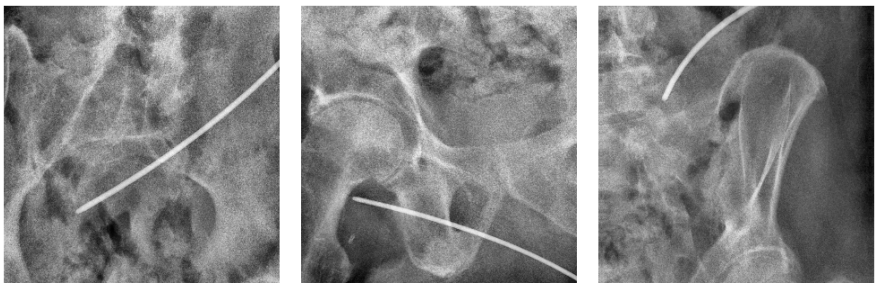

Figure 4: Example images from the generated training dataset with deformed simulated K-wires. Simulated B-spline based K-wire models with 3 control points were projected at varying poses on the augmented images of anatomy.

nearest interpolation for the masks to achieve the target image size, e.g., 224 × 224 pixels. Nearest interpolation was used for the masks in order to maintain their binary nature. The images were then saved in an 8-bit PNG format.

As shown in Fig. 2, to generate 2,000 images, we first select 80 real X-ray images, perform a ×25 randomized augmentation, and finally apply a 1:1 K-wire projection on each image, resulting in 2,000 images with K-wires and corresponding masks.

3.2 *Test* Dataset Generation

For our test dataset, we used an IRB-approved set of real patient images with K-wires on a pelvic background acquired with a Cios Spin (Siemens, Erlangen Germany). We selected the images with AP and AP-like views to get 20 images in total. Seven of those images had only one K-wire on them, while 13 had 2 K-wires. After we performed manual pixel-by-pixel segmentation of these images to get the ground truth location of the K-wire using Medical Imaging Interaction Toolkit (MITK), the images with multiple K-wires were split to create additional 26 images with single K-wires only, resulting in 33 total images in our test set as seen of Fig. 5. The images were downsampled to match the image size of the training dataset, prior to inference.

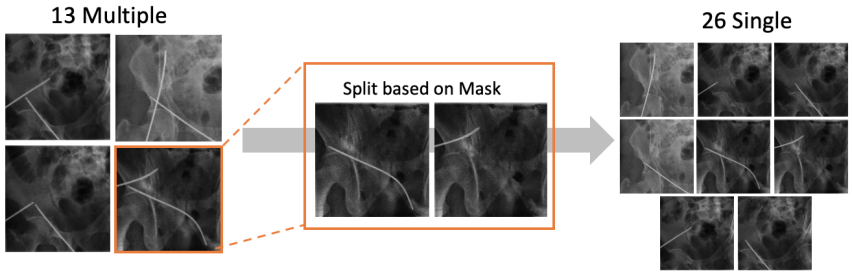

Figure 5: Flowchart depicting process to increase available test dataset size. Current analysis focuses on single K-wire detection. Test images obtained from an IRB-approved patient dataset were split based on manually annotated masks and inpainting into two images each containing a single K-wire.

3.3 Network Architecture Selection and Training

The goal of our project is a fast and accurate detection of K-wires in fluoroscopic images of the pelvis. We are thus interested in semantic segmentation of the image which outputs pixel by pixel binary classification.

Network Architecture. For this work, we have investigated U-Net and U-Net-like architectures. U-Net is a CNN found to be particularly successful in medical image analysis applications, where training set sizes are limited. We leveraged open sourced architectures established by [13] and [14] and adapted parameters to our task.

Network Backbone. The segmentation models we leveraged as our architecture foundation enabled selection of multiple backbones for the U-Net architecture. We initially tested and compared ResNet50, ResNet101, VGG16 backbones on a small dataset (250 image training, 50 image test). All backbones had comparable performance on predictions. Since Resnet101 resulted in longer inference and training time, we selected to move forward with Resnet50, since minimization of inference time supports our aim to increase the speed of K-wire detection. We plan to evaluate VGG16 in more depth as well in the next iteration of analysis.

Loss Function. We have tested the performance of various loss functions on the small dataset and performed metrics on AUC, recall, precision, and F1 Score. After evaluating performance on the test set, we observed that binary cross-entropy loss function (Eq. 1) performed similarly to weighted cross-entropy (Eq. 2) and both performed better than dice coefficient loss function (Eq. 3).

$$\mathrm{BCE}(y, \widehat{y}) = -(y \log(\widehat{y}) + (1 - y) \log(1 - \widehat{y})) \tag{1}$$

$$\mathrm{WCE}(y, \widehat{y}) = -(\beta y \log(\widehat{y}) + (1 - y) \log(1 - \widehat{y})) \tag{2}$$

$$\mathrm{DL}(y, \widehat{y}) = 1 - \frac{2y\widehat{y} + 1}{y + \widehat{y} + 1} \tag{3}$$

Pre-trained Model Weights. Transfer learning from non-medical images (e.g., ImageNet dataset) to medical images has been proven by other groups to be ultimately unsuccessful [15]. However, to confirm that this trend continued for our task and dataset, we tested network performance with loading the ImageNet pre-trained weights on a small 250 training data set run for 30 epochs. While loading ImageNet pre-trained weights resulted in shorter requried training times, training the network from scratch achieved comparable performance within 30 epochs.

The important note here is that employing weights pre-trained using ImageNet as our initialized weights requires conversion of a grayscale fluoroscopic image to RGB and the prediction will also be performed on the RGB image. Due to increase in color channels, this increases inference time, training time, and makes it harder to convert prediction output back to grayscale. We have therefore chosen not to use ImageNet weights as starting weights for our network.

4 Figures of Merit

4.1 Conventional CNN Prediction Metrics

These are the common metrics that are used for evaluation of semantic segmentation model performance and that we also chose for evaluation of the predictions for our project. TP stands for the number of true positive predictions, FP is false positives, FN is false negatives, while TPR and FPR are true positive rate and false positive rate, respectively.

$$\text{AUC} = \int \text{TPR(y-axis) vs. FPR(x-axis)} \tag{4}$$

$$\text{FPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{TPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

The objects that we are predicting on the image are only a small percentage of the whole image. Therefore, there is an inherently high percentage of TN predictions. This explains why we see most metric values jump quickly to 99%. We are ultimately most interested in the prediction of this last percent of pixels in the image, as these are the object of interest.

As we are interested in using these predictions for K-wire localization using multiple images, we are mostly interested in the TPR and FNR. The TP would most likely filter out when using multiple images, while a missing piece of K-wire due to high FN percentage would be more difficult to account for. Therefore, we primarily evaluate metrics based on the Recall score (Eq. 6) as improved recall score indicates lower numbers of FN. Since we have developed the network to favor FP, our Precision scores (Eq. 5) are always low and not as important as Recall.

4.2 Task-Specific Metrics

As the overarching goal of the project is to decrease the injuries due to K-wire misplacement, it is essential that the entire path of the K-wire is predicted accurately. We have therefore implemented 2 custom quantitative metrics for evaluation of the accuracy of the prediction. See Fig. 6 for the overview of the workflow of the metric calculation.

Hausdorff Distance to Centerline. This metric described the distance between the centerline of the prediction and the centerline of the ground truth image. It indicates how close the path of the predicted K-wire was to the path of the ground truth K-wire. To calculate this figure of merit, the prediction was first binarized with a threshold of 0.1, then dilated to increase the bright areas while minimizing the dark ones. We then identified connected

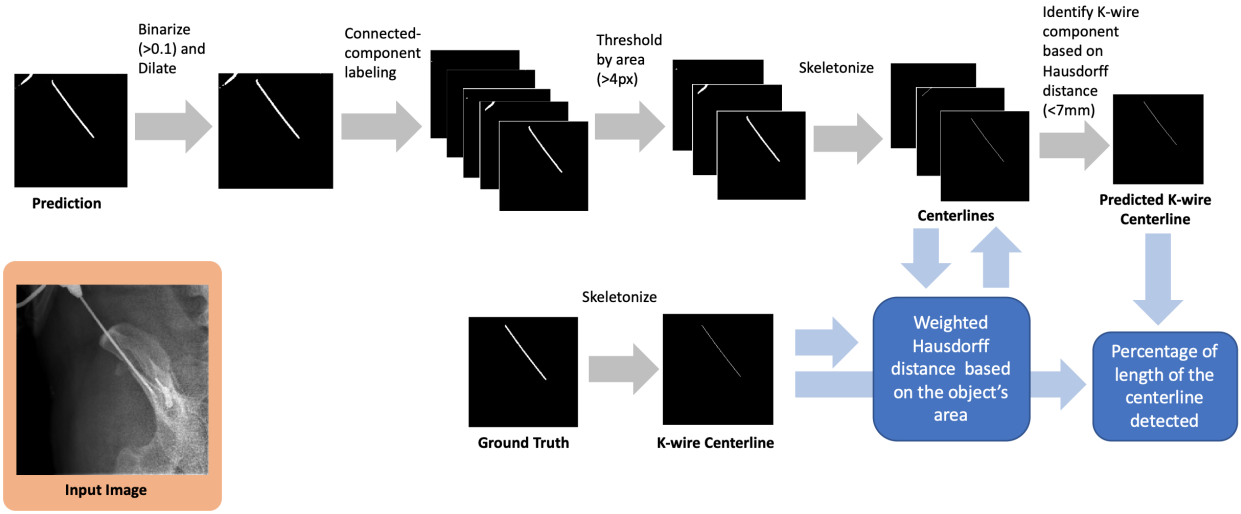

Figure 6: Workflow for the generation of task-specific performance metrics. Center-line comparison between prediction and ground truth are evaluated based on Hausdorff distance calculations. Using the measured Hausdorff distances, the percentage of the K-wire detected is calculated for detections that are < 7 mm away from the truth.

components on the image and thresholded components containing ¡4 pixels (outliers). Each of those were then skeletonized to obtain the centerlines. The ground truth was also skeletonized. For each skeletonized component of the prediction, the Hausdorff distance was calculated with the centerline of the ground truth. The weighted sum of all Hausdorff distances of the components was calculated based on the area of the component: the larger components had a larger weight, while smaller components had proportionally smaller weights.

Percentage of K-wire Length Detected. This metric describes the percentage of the length of the predicted K-wire compared to the actual length. It indicates how much of the length of the K-wire has been correctly predicted. To calculate this figure of merit, first the Hausdorff Distance to the centerline was computed. Only the components of the image that correspond actual K-wires – i.e., centerlines less than 7 mm away from the ground truth, were selected. The length of the component was then approximated as the length of the array of the coordinates of its centerline. The sum of all of the selected components' lengths was then calculated and compared to the expected length of the centerline to obtain the percent difference.

5 Results

As a primary aim for this project is to evaluate the success of our model's ability to successfully undergo transfer learning, our initial results focus on an analysis of how varying parameters impacts performance on real test images.

Training Time. We evaluated training time by training for 200 epochs using our 2,000 image training dataset. We observe on the validation loss plot in Fig. 7 that the model appears to begin to over fit after reaching the 80^{th} epoch.

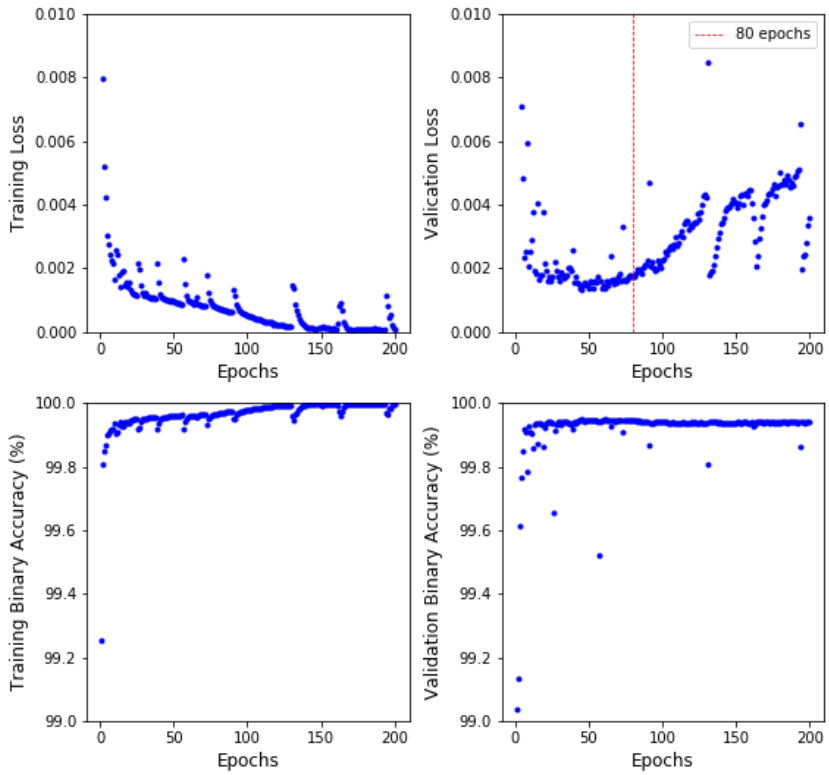

Figure 7: Training loss and accuracy as a function of training time. The top right plot, measuring validation loss over training time, shows that the model begins to overfit after the 80^{th} epoch.

Training Dataset Size. As medical images, even those without annotations, are very difficult to acquire and use, we had a significant interest in evaluating the necessary training population size needed to achieve our desired accuracy levels. Prior work using simulated training sets had in the range of 9,000 images [10]. However, as we were able to achieve light success with only 250 images, we decided to evaluate performance leveraging smaller training population sizes. As can be seen in Figs. 8, 9, and 10, as training population size increases, performance also tends to increase and converge.

Transfer Learning. Following training and validation on the simulated data set we evaluated the ability of the trained model to accurately detect K-wires in real pelvic images naturally containing K-wires. As seen in Fig. 11, the network (2,000 training set size, binary cross-entropy loss, 224×224 input image resolution) performs well on the real test set, showing overall success in our transfer learning approach. However, there are key failure regions for our network, in particular when the K-wire overlaps with the iliac crest. Here the contrast between the K-wire and background is reduced, which our network is not able to perform on. We hypothesize that adding additional noise could enable the model to perform better in these cases.

Other Parameters. In addition to training population size, we evaluated a number of other parameters including impact of image resolution, diversity of background image selection, and K-wire projection shape on the performance of our model. Rigid and Deformed

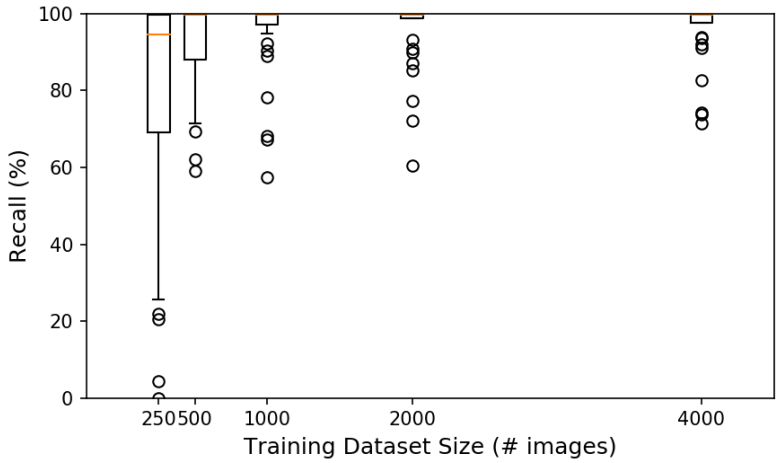

Figure 8: Recall as a function of training dataset size. Recall metrics are obtained by using the model weights from the 80^{th} epoch to perform predictions on the 33 images in the test dataset. Performance was observed to improve with increased training set size.

performed similarly (Fig. 12), the higher variation in background of the training dataset produced higher recall on the test dataset (Fig. 13), and the higher resolutions of the image produced higher recall score (Fig. 14).

Other Network Architectures. We performed comparison to a more complex U-Net variant U-Net++ [14], which redesigns the skip skip-connections within the original architecture. As can be observed on the Fig. 15 that shows the performance of U-Net found performance to be comparable, though inference time was doubled. Therefore, we feel that the original U-Net architecture is better suited out of the two for our task.

Inference Time. Inference time per image, leveraging the models trained using the U-Net architecture on both the 2,000 and 4,000 training datasets, yielded an inference time of 0.25 seconds per image on a workstation equipped with a Nvidia GeForce GTX 980 Ti. Using the U-Net++ architecture resulted in inference times of 0.5 seconds per image. The next phase of our analysis will delve further into establishing robust criteria to measure inference time as well as assess how this time can be brought down, such as implementing a more light weight model.

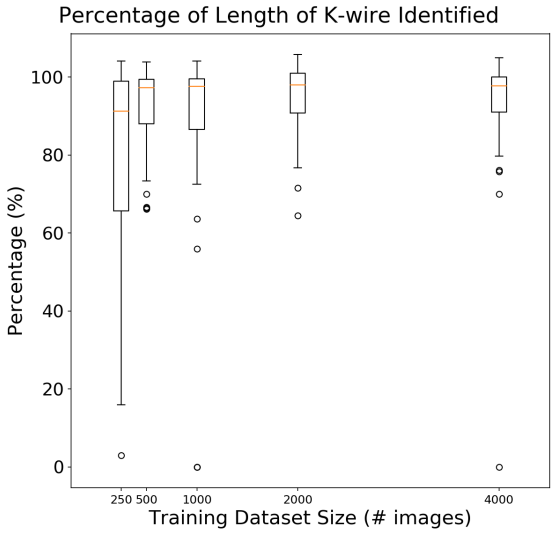

Figure 9: Percentage of K-wire length detected as a function of training dataset size. Detection accuracy increased with training dataset size, confirming trends for standard CNN metrics.

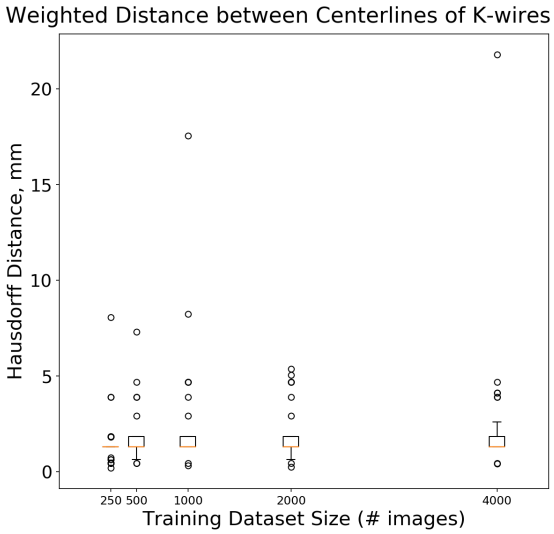

Figure 10: Hausdorff distance to centerline as a function of training dataset size. Differentiation between training set sizes is more limited indicating that for the portion of K-wire detected, the path of the K-wire is accurate.

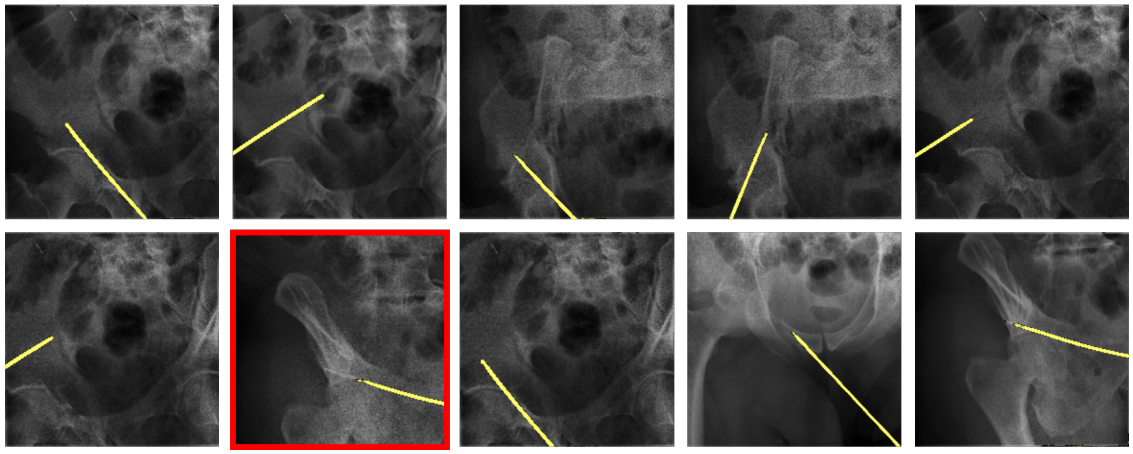

Figure 11: Example predictions on the test dataset. Images from test data set are overlaid with prediction results (yellow). Red box highlights a particular failure case caused by low contrast between the K-wire and background near the iliac crest.

6 Discussion

Through this work we have successfully demonstrated that a model developed using the U-Net architecture trained on a simulated dataset has a high prediction accuracy on real patient images. Given the success of the transfer learning, we believe a primary contribution from this work is our data generation workflow for the training dataset. We also defined key parameters yielding performance improvements for training the model: increase in background variability of training set, training dataset size, and potentially increasing image resolution.

Though we constrained our testing to AP-like views with single K-wires, our test dataset

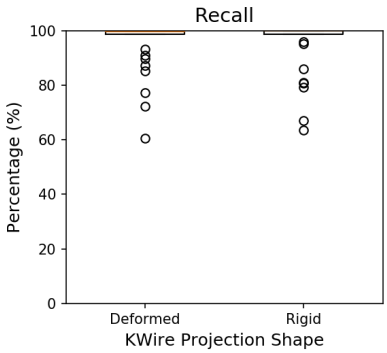

Figure 12: Training on deformed vs. ridid training datasets. Weights from the 80^{th} epoch are used for predictions. K-wire shape does not seem to have a big impact, indicating that the model is focusing more on image characteristics than overall object shape.

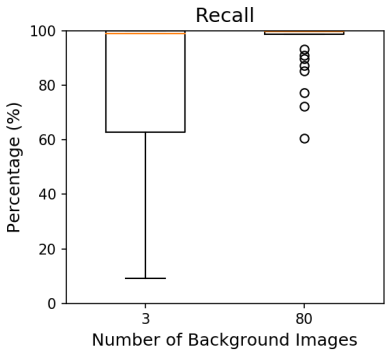

Figure 13: Training on dataset based on $N = 3$ vs. $N = 80$ background images. More input images, hence a greater diversity in backgrounds in the training set, yields improvement in model performance.

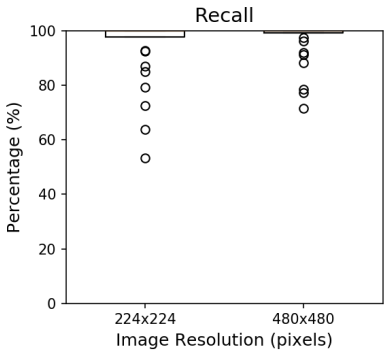

Figure 14: Training with input image resolution of 224×224 vs. 480×480. Input resolution does not appear to have a large impact on performance, although training time greatly increases with larger images and batch size has to be reduced to accommodate increased memory consumption.

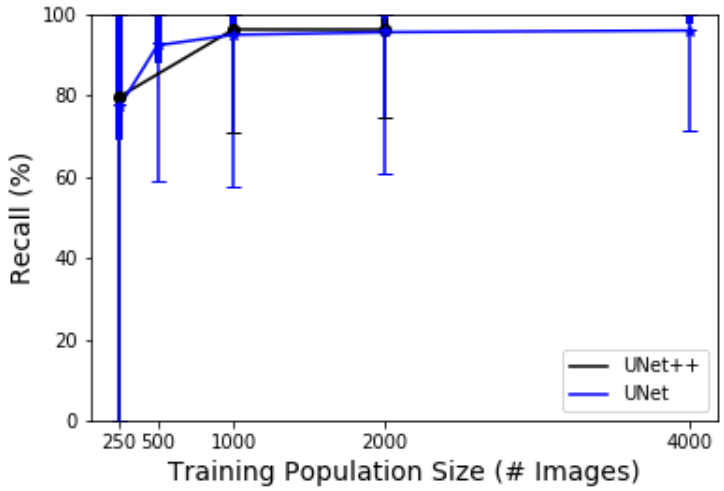

Figure 15: Comparison of U-Net and U-Net++ architectures. Weights from the 80^{th} epoch were used to perform prediction on the test dataset. Recall shows that models trained using either architecture yields similar performance.

was still noticeably visually different from the simulated dataset: it had higher noise, and images had lower overall intensity / contrast than those in the training dataset. Although the prediction had high accuracy for most images, accuracy tended to fail in the areas of

overlapping high intensity structures such as iliac crest, as seen in Fig. 11. The image highlighted by the red box depicts an example case of when the model fails to predict a significant portion of the tip of the K-wire. In the bright areas of the background, the intensity of the K-wire matched the intensity of the image, which could be the reason for the failure of the model to predict the K-wire. To address this, we may need to increase the noise applied to the images, to enable the network to learn to detect the shape even in instances of low contrast.

With regards to our task-specific evaluation metrics, while in most cases the task-specific metrics have been able to present an accurate picture of the quality of the prediction, we noticed some cases where the metrics failed to provide accurate representation of performance. As can be seen on Fig. 9, in many cases of the successful predictions, the percentage of the K-wire detected was calculated to be greater than 100%. This could be attributed to the method of the length calculation: it was approximated to as the relative length of the centerline array. Thus, if the centerline of the prediction had even small disturbances and appear noisy, that would be recorded as a greater length than if the line was smooth. One of the possible methods to correct for this would be to apply Savitzky-Golay filter to the centerline or fit a polynomial to a centerline to compute the length of the curve. Moreover, for some test images, the percentage of the length of the K-wire outputs is 0% even though upon visual inspection all or a significant portion of the K-wire was predicted. The reason for such discrepancy is the threshold method for calculation of the metric: only the components which centerlines are less than 7mm away from the ground truth K-wire centerline were included into the percentage. Some of the images, however, contained tool tips and noise, which extended the ends K-wires. These extensions were still part of the identified K-wire component but had a large Hausdorff distance associated which was extrapolated to the whole component. More work is therefore needed to improve the accuracy of this metric. These issues with the predictions that were exposed by this metric, however, would unlikely to significantly affect the accuracy of the the goal of K-wire localization. These deficiencies in prediction would most likely disappear if multiple images are used for determining the shape of the kK-wire.

A surprising observation during our experiments was that our model, trained exclusively on images with single K-wires, was able to also perform well on test images with multiple K-wires. Results are not provided in this paper as further analysis is necessary.

7 Future Work

While significant progress has been made towards a robust K-wire detection process leveraging deep learning, there is need for future work to achieve our accuracy and speed goals.

The near term goal of the project would be to improve upon the transfer learning process. This could be achieved through acquiring more real images with K-wires to use as part of the training set, as implemented in related works [10]. Another way would be to add new parameters or expand the current range of the parameters for data augmentation. As discussed previously, the predictions were most likely to be poor in the areas of high image intensity. The training dataset can thus be expanded to include images with K-wires that cross through high-intensity regions of the image to potentially improve the prediction.

Further analysis is needed to evaluate the impact of input resolution during training on model performance. Although it appeared to improve the accuracy of the prediction (Fig. 14), it also results in longer training and inference times. We would need to determine whether improvements to performance are statistically significant in order to warrant the trade-off.

U-Net is a state of the art network with many variations that have been developed since it was first introduced in 2015 [16]. While we have tested U-Net++ architecture as part of the project [14], we also plan on testing U-Net Light, which was described to have a fast inference time without the loss of accuracy of the prediction [10]. While we have achieved good results for semantic segmentation, the application of the K-wire detection will realistically involve multiple K-wires on the same image. We are therefore aiming to employ object detection. To achieve this goal, we will be assessing the performance of U-Net and multi-class predictions, as well as exploring additional object detection architectures such as Mask-RCNN.

Lastly, we aim to implement 3D localization for both single and multiple K-wires. Once it is achieved, the figures of merit such as tip's position accuracy and orientation can be calculated and compared to the prior work.

8 Conclusion

Our work lays the foundation for rapid 3D localization of K-wires in fluoroscopic images of the pelvis using deep learning. The approach offers fast and accurate detection of surgical instrumentation in 2D radiographic images that can help address the limited capture range and slow runtime of existing methods for 3D-2D registration through robust initialization. Through the development of a large simulated dataset of x-rays of the pelvis containing K-wires, we remove the dependency on access to large, manually annotated datasets for the development of deep learning based tools. We evaluate the success of our simulated dataset in enabling transfer learning by implementing a U-Net architecture, training a model on our dataset, and leveraging the output weights to perform predictions on an IRB-approved, real patient dataset containing K-wires. The inference time and error metrics provide a proof of concept that a deep learning based workflow can provide both high accuracy as well as speed for K-wire detection in surgical navigation, which would enable a reduction in injury while also limiting raditaiton exposure. Through our analysis, we establish baseline metrics for K-wire identification tasks upon which further analysis can be performed.

9 Deliverables

Minimum: Train Model on Simulations

- Simulated K-wire images.
- Trained and validated CNN to detect K-wire in 2D radiographs.
- Figures of merit for simulated data results.
- Documented code leveraged for model development and training on simulated data.

Expected: Evaluate Model on Real Data

- Establish "real image" data set.
- Transfer Learning from simulated data to real data.
- Figures of merit on "real image" data set.
- Code documentation.

Maximum: 3D Localization and Guidance

- Evaluations of existing stereo x-ray images to 3D localization implementations.
- Design of 3D localization algorithm.
- Evaluation of dose/x-ray protocols.
- Documentation of code and protocols.
- Conference submission.

We have completed both minimum and expected deliverables. Due to lab closures related to COVID-19, our initial plan on obtaining and using cadaver images with real K-wires for tuning training and testing had to be changed. We therefore reworked the plan to complete a more thorough evaluation of the model and test different architectures. Later, however, our mentor Dr. Ali Uneri has been able to locate images that we could use for our test dataset, which changed our schedule again. Both of the team members, however, are planning to continue working on the project and completing the maximum deliverable.

10 Project Management

While our team worked collaboratively on all parts of the project, we had assigned responsibility ownership of different parts of the project to different members to maintain accountability.

Irina Bataeva: K-wire simulation, test set segmentation, model evaluation using task-specific metrics.
Kinjal Shah: Training dataset augmentation, training model design and development, model evaluation using standard CNN metrics.

Meeting Cadence:

- Weekly meetings as a group every Sunday, Wednesday
- Weekly meetings with Dr. Ali Uneri on Fridays at 3:30pm
- Slack channel available for ad-hoc communication

Code and documentation is maintained on GitLab at https://git.lcsr.jhu.edu/auneri1/kwiredetection

11 Acknowledgements

We would like to thank our mentor, Dr. Ali Uneri, for all of his guidance and support throughout this project. We would also like to thank Dr. Jeffrey Siewerdsen and the I-STAR lab for providing access to images and computational resources, Professor Taylor and Baichuan for teaching us the tools needed to execute a successful project, and the authors of "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation", for a well documented code repository upon which we could build our work.

References

[1] L. T. Buller, M. J. Best, and S. M. Quinnan, "A nationwide analysis of pelvic ring fractures," *Geriatric Orthopaedic Surgery & Rehabilitation*, vol. 7, no. 1, p. 9–17, Mar 2015. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/2151458515616250

[2] J. Mcmaster, "Pelvic ring fractures: assessment, associated injuries, and acute management," *Oxford Medicine Online*, 2011.

[3] C. C. L. Vu, R. P. Runner, W. M. Reisman, and M. L. Schenker, "The frail fail: Increased mortality and post-operative complications in orthopaedic trauma patients," *Injury*, Aug 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020138317305466

[4] D. Schweitzer, A. Zylberberg, M. Córdova, and J. Gonzalez, "Closed reduction and iliosacral percutaneous fixation of unstable pelvic ring fractures," *Injury*, vol. 39, no. 8, pp. 869 – 874, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020138308001708

[5] J. Görres, A. Uneri, M. Jacobson, B. Ramsay, T. Silva, M. Ketcha, R. Han, A. Manbachi, S. Vogt, G. Kleinszig, J.-P. Wolinsky, G. Osgood, and J. Siewerdsen, "Planning, guidance, and quality assurance of pelvic screw placement using deformable image registration," *Physics in Medicine and Biology*, vol. 62, 10 2017.

[6] R. Rampersaud, D. Simon, and K. Foley, "Accuracy requirements for image-guided spinal pedicle screw placement," *Spine*, vol. 26, pp. 352–9, 03 2001.

[7] V. Dzupa, J. Chmelova, P. Obruba, P. Wendsche, and P. Simko, "Multicentric study of patients with pelvic injury: basic analysis of the study group," *Acta Chirurgiae Orthopaedicae et Traumatologiae Cechoslovaca*, vol. 76, no. 5, p. 404–409, Sep 2009. [Online]. Available: https://europepmc.org/article/med/19912705

[8] G. Poole, E. Ward, J. Griswold, F. Muakkassa, and H. Hsu, "Complications of pelvic fractures from blunt trauma," *The American surgeon*, vol. 58, no. 4, p. 225—231, April 1992. [Online]. Available: http://europepmc.org/abstract/MED/1586080

[9] J. P. Zwingmann, O. P. Hauschild, G. P. Bode, N. P. Südkamp, and H. P. Schmal, "Malposition and revision rates of different imaging modalities for percutaneous iliosacral

screw fixation following pelvic fractures: a systematic review and meta-analysis," *Archives of Orthopaedic and Trauma Surgery*, vol. 133, no. 9, p. 1257–1265, Aug 2013. [Online]. Available: https://link.springer.com/article/10.1007/s00402-013-1788-4

[10] M. Gherardini, E. Mazomenos, A. Menciassi, and D. Stoyanov, "Catheter segmentation in x-ray fluoroscopy using synthetic data and transfer learning with light u-nets," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105420, 02 2020.

[11] M. G. Wagner, P. Laeseke, and M. A. Speidel, "Deep learning based guidewire segmentation in x-ray images," *Medical Imaging 2019: Physics of Medical Imaging*, Jan 2019.

[12] C. Shorten and T. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, 12 2019.

[13] P. Yakubovskiy, "Segmentation models," https://github.com/qubvel/segmentation_models, 2019.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, 2019.

[15] Raghu, Maithra, Zhang, Kleinberg, Jon, Bengio, and Samy, "Transfusion: Understanding transfer learning for medical imaging," *arXiv.org*, Oct 2019. [Online]. Available: https://arxiv.org/abs/1902.07208

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science*, vol. 9351, 10 2015, pp. 234–241.